Available online at www.sciencedirect.com

ScienceDirect

Journal of Biomedical Informatics 40 (2007) 649–660

Journal of Biomedical Informatics

www.elsevier.com/locate/yjbin

Discovery and inclusion of SOFA score episodes in mortality prediction

Tudor Toma [a,*], Ameen Abu-Hanna [a], Robert-Jan Bosman [b]

[a] *Department of Medical Informatics, Academic Medical Center, Universiteit van Amsterdam, P.O. Box 22700, 1100 DE Amsterdam, The Netherlands*

[b] *Department of Intensive Care, Onze Lieve Vrouwe Gasthuis, Amsterdam, The Netherlands*

Received 19 September 2006

Available online 31 March 2007

Abstract

Predicting the survival status of Intensive Care patients at the end of their hospital stay is useful for various clinical and organizational tasks. Current models for predicting mortality use logistic regression models that rely solely on data collected during the first 24 h of patient admission. These models do not exploit information contained in daily organ failure scores which nowadays are being routinely collected in many Intensive Care Units. We propose a novel method for mortality prediction that, in addition to admission-related data, takes advantage of daily data as well. The method is characterized by the data-driven discovery of temporal patterns, called episodes, of the organ failure scores and by embedding them in the familiar logistic regression framework for prediction. Our method results in a set of D logistic regression models, one for each of the first D days of Intensive Care Unit stay. A model for day $d \leqslant D$ is trained on the patient subpopulation that stayed at least d days in the Intensive Care Unit and predicts the probability of death at the end of hospital stay for such patients. We implemented our method, with a specific form of episodes, called aligned episodes, on a large dataset of Intensive Care Unit patients for the first 5 days of stay ($D = 5$) in the unit. We compared our models with ones that were developed on the same patient subpopulations but which did not use the episodes. The new models show improved performance on each of the five days. They also provide insight in the effect of the various selected episodes on mortality.

Keywords: Temporal patterns; Pattern discovery; Logistic regression; Prognosis; Intensive Care; Organ dysfunction scores

1. Introduction

Reliable clinical predictions of future events are useful in supporting various clinical and managerial tasks such as optimizing workload schedules, and assessing quality of care [1]. In the Intensive Care (IC), where patient survival forms an important indicator for the effectiveness of care, prognostic models to predict the probability of the survival status of patients upon discharge are commonplace. The most common application of prognostic models is in comparative audit among IC units (ICUs). In this setting a prognostic model for predicting mortality of ICU patients at their discharge from hospital is developed based on retrospectively collected data of all participating ICUs. These data depict the demographics and severity of illness of the patient during the first 24 h of patient admission to the ICU. The model is then used to predict mortality of new patients based on prospectively collected data to each ICU. Discrepancies between the predicted and actual mortality, measured in terms of the *Standardized Mortality Ratio*, is used to rank the performance of the various ICUs. Because the predictions adjust for the severity of illness of the patients, which is called case-mix adjustment, discrepancies between the predicted and the actual mortality is assumed to be attributed to the quality of delivered care. Regular application of comparative auditing contributes to assessing and improving ICU quality

* Corresponding author. Fax: +31 20 6919840.
E-mail address: t.toma@amc.uva.nl (T. Toma).

1532-0464/$ - see front matter
doi:10.1016/j.jbi.2007.03.007

of care. Current ICU prognostic models such as the Simplified Acute Physiology Score (SAPS-II) [2] and the Acute Physiology And Chronic Health Evaluation (APACHE) [3] summarize admission data in terms of severity-of-illness-scores, which take integer values where a higher score indicates a more severe health condition of the patient. These scores are then used as covariates in a logistic regression model (see Appendix) to predict the probability of the survival status of a patient at discharge from the hospital.

Intensive Care is, however, a dynamic medical environment where patients' health status can change rapidly in either direction. Capturing and analyzing such dynamics can provide better insight in patients' health status over time. Recently a growing number of ICUs have started collecting, in addition to the static data, Sequential Organ Failure Assessment (SOFA) [4] scores to assess the incidence of organ dysfunction. A SOFA score is a daily quantitative assessment of the patient's organ dysfunction during ICU stay. The SOFA score is an integer ranging from 0 to 24, where a greater value corresponds to worse organ function derangement. Its value is the sum of 6 individual organ system scores, each ranging from 0 to 4. The process of documenting SOFA scores generates much temporal information which calls upon computer aided tools to analyze it as it has been demonstrated that doctors have difficulty in interpreting temporal information [5]. Although not specifically developed to assist mortality prediction, it was natural to investigate the predictive value of SOFA scores. There are many studies demonstrating that non-survivors have a significantly higher mean SOFA score than survivors [6]. However, such observations do not provide a model for predicting mortality of a given patient on a given day. Almost all current work on SOFA-based mortality prediction rely solely on simple pre-specified summaries, such as the mean, in the patient's SOFA score sequence known prior to the day upon which prediction is to be made. The prognostic merits of SOFA sequences that preserve the temporal evolution of organ functioning over time, are unclear and have been hardly investigated.

Prognostic models adjusting for the raw SOFA scores of patients are not meant to be used in the same way as the static ones for comparative auditing among ICUs. This is because a SOFA score is influenced by treatment. For example, the SOFA score of a severely ill patient at admission in a well performing ICU could be the same as that of a relatively healthier patient that is receiving sub-optimal therapy. Mortality predictions of both patients will be the same, blurring the effect of treatment. However, SOFA-based prognostic models provide insight into the dynamics of organ failures and their relation to mortality, and they are useful for other tasks. First, they can be used to compare performance of the same ICU over time. Second, accurate probabilities can help physicians to proactively decide on intensifying interventions for patients with worsening prognosis. Third, predictions can be used to help managers to better plan capacity e.g. of nursing services. The purpose of this work, however, concerns the development and performance of the temporal predictive models and not their clinical and managerial application.

The objective of this paper is to propose a method for the representation, selection and inclusion of SOFA score patterns, called episodes, in mortality prediction. These episodes are meant to be used *in addition* to the SAPS in order to be able to investigate their added predictive value. The application of this method results in a set of D models, called temporal models, $\{M_d\}, 1 \leqslant d \leqslant D$, one for each of the first D days of ICU stay. At day $d \leqslant D$ of stay, model M_d predicts for the patients that stayed at least d days in the ICU, the probability of death at the end of their hospital stay.

In a nutshell, our approach is based on the discovery of frequent qualitative episodes of SOFA scores. This is done for each day on which prediction of the eventual vital status is to be made. These non-prespecified episodes are easily interpreted as the temporal evolution of patients' health status during their stay in the ICU. To use them in prediction, the frequent episodes are represented as indicator variables for possible inclusion in a logistic regression model predicting the probability of mortality. In this way episode discovery is integrated into the current solid logistic regression framework for prediction. This also has the advantage of providing a fair comparison with the existing static logistic regression approaches allowing to make a judgement about the added value of SOFA episodes in prediction. The method was applied to data from an adult ICU of a teaching hospital. We obtained temporal models for mortality prediction in each of the first five days since admission. For comparison we also developed similar five models that used only the static data obtained on the day of admission. The resulting temporal models reveal the quantitative and qualitative associations between the selected episodes and mortality. Model validation shows that the obtained temporal models provide better predictions than the static ones on the validation set.

The paper is organized as follows: Section 2 describes the data and our method in terms of a workflow consisting of four tasks. We present our results in Section 3. Section 4 discusses our results in the context of related work and concludes the paper.

2. Materials and methods

The proposed method for the development and evaluation of the prognostic models is conveniently described in engineering terms as consisting of the following tasks: Data preprocessing; Discovery of frequent episodes from SOFA scores; Development of prognostic models using frequent episodes; and Evaluation of the resulting models. These tasks are interrelated in the workflow model shown in Fig. 1 and are described below. The description of these tasks is illustrated by a real world application in the IC.

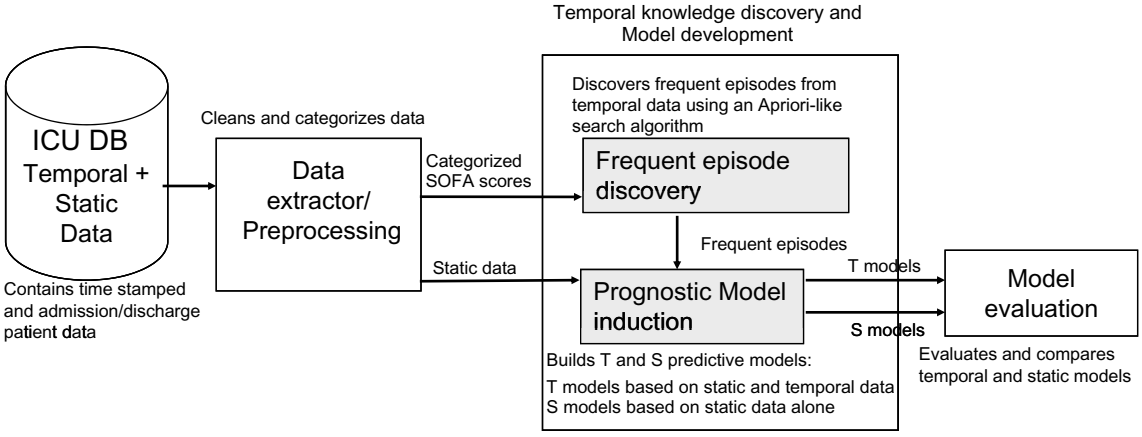

Fig. 1. Workflow showing the four primary tasks in our approach.

2.1. Data and data preprocessing

The application considered in this paper concerns ICU patient data from the OLVG, a teaching hospital in Amsterdam. The data was collected between July 1998 and August 2005 and includes all consecutive 6865 ICU admissions in that period. The data contains demographic information such as age and sex, and physiological and laboratory information collected during the first 24 h of admission. From this information one calculates the commonly used severity-of-illness-score SAPS-II (hereafter SAPS in short) for all patients, without applying the SAPS exclusion criteria, because the SAPS is not the focus in this study. In addition, contrary to the SAPS exclusion criteria, in case patients were readmitted to the ICU (which occurred in 5.4% of the cases) only their last readmissions were kept due to their relevance to predicting the outcome. The SAPS scores, and the models derived solely from them will be referred to as static data and static models, respectively. In addition to the static data, the database includes daily SOFA scores. For each patient there corresponds a sequence of SOFA score values, one value for each day of stay. To exemplify, a patient's ICU stay of three days may correspond to the SOFA sequence 13–11–10 which indicates a slight improvement in organ functioning during this period. The SOFA scores, and the models derived from them (including also static data) will be referred to as temporal data and temporal models, respectively.

There were some artifacts induced by data collection. Missing SOFA score values, amounting to 79 cases, were imputed by the mean value of their adjacent SOFA scores. Cases with at least 2 consecutive values missing, amounting to 12 cases, were not considered in the analysis. After removing earlier readmissions and correcting for these artifacts, 6276 of the original 6865 admissions were retained for further analysis. Table 1 characterizes the survivor and non-survivor patients. There are 5587 survivors and 689 non-survivors (corresponding to 11% hospital mortality). Hospital mortality refers to deaths in the hospital during, or after, stay in the ICU. Admission type describes the reason for admitting the patient: due to medical reasons, or due to a prior surgery (urgent or planned) necessitating subsequent ICU stay. Length of stay (LOS) denotes the total number of days a patient stayed in the ICU until discharge.

Table 1
Static data descriptive statistics

	Survivors	Non-survivors
N	5587	689
Admission type (%)		
Medical	16.4	70
Urgent	9.4	15
Planned	74.2	15
Mean age	64	68
±SD (years)	±14.3	±14.6
Sex		
Male/female (%)	66.7/33.3	59/41
Median LOS	0.93	2.18

Every patient receives a SOFA score at admission, at 6:00 am of every morning during ICU stay, and at discharge. This means that the number of LOS in days may not exactly correspond to the number of SOFA scores assigned in the patient's record. For a patient that is admitted just a couple of hours before 6:00 am will have a SOFA score at admission and another one at 6:00 am. Based on medical expert knowledge it was decided that only periods of at least 6 h between admission and 6:00 am, and between 6:00 am and discharge "deserve" a SOFA score. For simplicity of analysis and presentation we will consider from now on this adjusted number of SOFA scores as representing the number of days that a patient has stayed in the ICU, disregarding the exact number of hours. This means that patients with one SOFA score are considered as staying one day in the ICU even if they have stayed only for half a day.

Table 2
Summary statistics of SAPS and SOFA scores in patients staying at least one day (1 SOFA score) and the patients staying at least 5 days (5 SOFA scores)

	Survivors	Non-survivors
Mean SOFA ± SD		
LOS ⩾ 1	7.3 ± 1.8	10.4 ± 3.2
LOS ⩾ 5	8.5 ± 2.6	10 ± 2.9
Mean SAPS ± SD		
LOS ⩾ 1	31 ± 12	60 ± 20
LOS ⩾ 5	49 ± 15	57 ± 17
N		
LOS ⩾ 1	5587	689
LOS ⩾ 5	444	144

Table 2 shows the mean SOFA and SAPS scores for patients staying at least one day (all patients) and for the subpopulation that stayed at least 5 days (i.e. received at least five SOFA scores). Note the higher mean of SAPS and SOFA scores in non-survivors than survivors. Moreover, for patients staying longer in ICU (at least five SOFA values) the mortality is higher than for the sample of all patients (24 vs. 11%). This provides some evidence to the theoretical utility of SOFA scores in prediction. These summaries however do not help quantify the association between SOFA sequences and probability of death nor whether the SOFA scores have added value beyond the information already residing in the SAPS.

The next step in preprocessing the data consists of categorizing the SOFA scores in a few qualitative categories. This data reduction is necessary for the next task of episode discovery. Instead of a large number of low frequency episodes in the raw SOFA score sequences, one obtains markedly fewer but much more frequent episodes of the categorized SOFA sequences. It also has the advantage of making the categorized sequences easily interpretable by a human analyst. In our application we have chosen for three categories: `LOW (L)`, `MEDIUM (M)`, and `HIGH (H)` based on the 3-quantile grouping of SOFA scores. Quantiles divide ordered data into equally sized data subsets. In the sequel we will refer to the qualitative SOFA score as qSOFA. The 3-quantile division of the SOFA resulted in the following categorization rules:

(1) if SOFA score $\in\{0,\ldots,6\}$ then it is coded as `LOW (L)`
(2) if SOFA score $\in\{7,8\}$ then it is coded as `MEDIUM (M)`
(3) if SOFA score $\in\{9,\ldots,24\}$ then it is coded as `HIGH (H)`.

For example the SOFA sequence 12–7–10–8–5 will be recoded as `HMHML` in both training and test sets.

2.2. Frequent episodes

We define an episode as a sequence of consecutive qSOFA scores. This is in line with the definition of a serial episode in [7]. The qualitative SOFA values are vertical (contemporaneous) data abstractions [8] or State Temporal Abstractions [9] where an interval of values is abstracted in a new data point in our case a SOFA category.

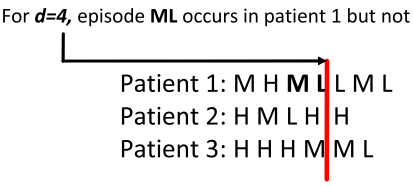

Fig. 2. Episode aligned to prediction day $d = 4$.

In this paper we investigate a special type of episodes called aligned episodes (Fig. 2) in which the last qSOFA value in the episode belongs to the day at which prediction is to be made. For example, if the episode `ML` is said to be aligned to the fourth day of stay it means that the `L` qSOFA score was obtained on the fourth day and the `M` on the third day. Such episodes can be used to make predictions at discharge time (e.g. which is unknown and can be for example on day 12) using information available until prediction time (e.g. here in day 4). The choice for episode alignment is motivated by our hypothesis that the days closest to prediction time are more relevant to the outcome than those earlier in the score evolution. This Markovian-like hypothesis is endorsed by the clinician involved in this study. We focus on episodes that are frequent in the data. This will greatly reduce the computational burden and boost the stability of the prognostic model variable selection process described shortly. For the development of models that are meant to predict mortality of all patients, and not e.g. for a subgroup of patients at very high risk, high frequency patterns are likely to be the most dominant in the model.

Algorithm 1. Frequent aligned episodes discovery algorithm for day d

- d - integer denoting the day at which prediction is to be made
- *Elm* - set of all elementary episodes, in this paper {`L`, `M`, `H`}

1: $PATd \Leftarrow$ set of patients who stayed at least d days in ICU
2: $E \Leftarrow \{\}$/* Initialize E. It will include all frequent aligned episodes */
3: $S_1 \Leftarrow$ set of all frequent elementary episodes in $PATd$ when aligned to d
4: $j \Leftarrow 1$
5: **while** not empty (S_j) **do**
6: $E \Leftarrow E \cup S_j$/* Add to E the frequent episodes of length j*/
7: $SB \Leftarrow$ Extend each element in S_j backwards with one element from *Elm*
8: $j = j + 1$
9: $S_j \Leftarrow$ set of all episodes in SB that are frequent in $PATd$ when aligned to d
10: **end while**

We developed a discovery algorithm, implemented in JAVA, that searches for frequent aligned episodes for any given day at which prediction is to be made. An episode aligned to day d is said to be frequent when its frequency rate in the subset of patients staying at least d days exceeds a pre-specified threshold (e.g. 5% of cases) referred to as minimum support rate. A pseudocode of the algorithm, which is a specialization of that proposed in [7], appears in Algorithm 1. To illustrate our algorithm, suppose that we are to find the frequent episodes for the third day of stay ($d = 3$). We must only consider the patients that stayed at least 3 days in the ICU because the outcome is already known for the patients who died or left the ICU prior to the third day. The search algorithm starts by generating all episodes of length 1—these will be `L`, `M` and `H`. For each episode it will count how often it occurred in the patients' qSOFA sequence on the third day. It will retain only those episodes occurring frequently enough. Suppose only `L` was retained. The process is repeated, extending the retained episode *backwards* by one element at each step resulting in `LL`, `ML`, and `HL`. Each of these episodes will be matched against the patients' qSOFA scores at the second and third day of stay. The process is repeated by retaining only the frequent episodes of length 2 and extending them backward until no more frequent episodes are found. This process is based on the A-priori property [10] which, in our case, states that a length-$k+1$ episode cannot be frequent if its length-k sub-episode is infrequent. In our application we have set the minimum support rate at 5%. The discovered frequent episodes can themselves be of interest to the clinician as they can enhance the insight in the patient population. However, they do not include any association with mortality yet. Below we discuss how to further put these episodes into use in the task of mortality prediction.

2.3. Model fitting strategy

For each day at which a prediction is to be made a separate model will be developed. In this work we will develop a model for each day of the first 5 days of ICU stay: $M_d, 1 \leqslant d \leqslant 5$. For M_d two components are required: the data from the training set corresponding to the first d days of stay (for patients that stayed at least d days), and the set of frequent episodes that were discovered for that day as described in the previous subsection. The inclusion of static data (i.e. SAPS) in the models that use frequent episodes is motivated by the fact that this allows us to assess the added prognostic value of temporal information in the context of the static models (SAPS only).

The model of choice in our method is logistic regression. This has the advantage of integrating temporal information in the already existent framework for building static prognostic models. Besides, the coefficients of the logistic model have an intuitive meaning (see Appendix for a short description of the logistic regression model). In order to show the added value of the episodes, the temporal models will always include the severity-of-illness-score SAPS as one of the covariates. To include the frequent episodes in the model we will code them as binary indicator variables to be considered as candidate dummy variables in the logistic regression model. A dummy variable describes whether the episode it represents is present (value 1) or not (value 0) in a patient's qSOFA sequence aligned to the day when prediction is required. For example, consider a patient with the first five days qSOFA sequence `LMHHM`. When developing the model M_5 the value of the dummy variables of the frequent episodes `HM` and `HHM` for this patient will be 1. However they will be 0 for a patient with a sequence beginning with `MHHMM`. For the frequent episode `M` the value of the dummy variable will be 1 for both patients.

Often the size of the set of all possible models created based on the frequent episodes belonging to any M_d, beyond the first day, is quite large. This would compromise the validity of a model created by exhaustive search because of the increasing likelihood of selecting sub-optimal episodes. For example if for $d = 5$ a set of 19 frequent episodes were discovered, the number of all possible models created with these will be $2^{19} \simeq 5.2 \times 10^5$. This problem would only become more difficult by using all possible episodes instead of only the frequent ones just for $d = 5$ a number of 363 potential episodes being available. Beyond the induced computational burden the bigger the search space the higher the likelihood to, by chance, select in the models variables that are not significant (Type I error) generating unstable models due to overfitting.

On the other hand not all frequent episodes are expected to be important in a model. For each M_d, a model fitting strategy, not suffering of the drawbacks of an exhaustive search, should be applied for selecting the best covariates by searching for the most important dummy variables in the space of all candidate dummy variables belonging to day d. One can use stepwise approaches for including or excluding covariates, one by one, and select the "best" model. The common practice of using significance testing based on p-values for model selection is not appropriate because the obtained p-values are meant to be used only for a pre-specified model [11]. The problem is only exacerbated when there is a large set of candidate covariates.

In our approach we avoid using significance testing for model selection and propose using an information-theoretic measure. In particular we will use Akaike's Information Criterion (AIC) [12] to evaluate the models. The AIC is defined as $-2\log L(\theta) + 2k$ where $L(\theta)$ is the maximized likelihood [13] of the model and k is the number of free parameters in the model. Hence the AIC trades off predictive performance for parsimony by penalizing for the number of variables included in the model. Parsimony

is important in order to avoid overfitting. To select the subset of the covariates resulting in the best model (i.e. with minimum AIC) it is common to fit a model including all candidate covariates and then to eliminate the least predictive covariate—the one associated with the highest AIC—one by one until the AIC cannot be decreased.

Applying this model fitting strategy has however a disadvantage, which is present in our case, when there is strong *collinearity* between the covariates. Collinearity is a situation in which at least one of the covariates can be predicted well from the other covariates [11]. An example of an obvious type of collinearity occurs when, for any given day d, one episode *logically entails* another, e.g. `MH` logically entails `H` (recall that episodes are aligned and `MH` implies that the patient had a qSOFA score of `H` on day d). During model selection the covariates compete for a place in the model. When there are collinear covariates, stepwise selection may make an arbitrary choice between them. This might result in biased models by omission of predictive variables from the model. In addition, the interpretation of the model's coefficients (see Appendix) will not be straightforward because it is based on the idea of studying a change of a covariate in isolation. But if covariates are collinear it is not possible to change only one covariate without affecting the other collinear covariates.

To deal with collinearity we will not include at once all candidate covariates of episodes of various lengths in the stepwise method: first only dummies of episodes of length 1 are included. Only those that survive the AIC based backward selection procedure will be included with dummies of episodes of length 2, etc. It makes sense to start with the most recent data prior to the day at which prediction is to be made, and include dummies of longer episodes incrementally. This "Markovian-inspired" choice infuses background knowledge in the process.

In addition, we do not allow for the type of collinearity in which an episode is logically entailed by any other. Whenever there is a set of logically entailed covariates in a model we will search for the best model, in terms of AIC, having no logically entailed covariates. For example, if the model includes, among others, `H`, `LH` and `MH` as covariates, two models will be assessed: one with `H` (without `LH`, `MH` because they will otherwise logically entail `H`) and one with `LH` and `MH` (without `H`). The model with the lower AIC will be retained. Algorithm 2 shows the pseudocode for the model fitting strategy. Each episode in E has an associated element in B which is an indicator variable. The indicator variable takes the value 1 when the episode, aligned to the day of prediction, occurs in a patient's sequence. The indicator variables are considered as dummy covariates for possible inclusion in the logistic regression models. The function *binaryToEpisode*() takes a dummy variable and returns the episode it represents. For predicting mortality at day d, all dummy variables of frequent episodes of length 1 to d will, incrementally, be considered.

Algorithm 2 Model fitting strategy for day d

- d—day at which hospital mortality prediction is required
- E—set of frequent episodes - length 1 to d
- B—set of indicator variables for episodes in E, used as candidate dummy covariates
- *binaryToEpisode*: B → E, returns the episode associated with a dummy candidate covariate

1: $Cov \Leftarrow \{SAPS\}$ /*initialize the covariates set with the SAPS*/
2: **for** $j = 1$ to d **do**
3: $Cands \Leftarrow Cov \cup \{b \in B | \textbf{length}(\textbf{binaryToEpisode}(b)) = j\}$ /* add to the covariates dummies associated with episodes of length j*/
4: $fullModel \Leftarrow$ **fitLogReg**($Cands$) /* the full logistic regression model*/
5: $Cov \Leftarrow$ **variableSelection**($fullModel$,$type =$ "$back$", $method =$ "AIC") /*stepwise backward variable selection using the AIC*/
6: **if collinears**(Cov)
7: $Cov \Leftarrow$ **eliminateCollinears**(Cov) /* eliminate logical entailment collinearity between covariates if it exists*/
8: **end if**
9: **end for**
10: $Model \Leftarrow$ **fitLogReg**(Cov) /* the final prognostic model is fitted using the selected covariates*/

We implemented the algorithm in S-Plus statistical environment where we used the `MASS` library with its `stepAIC` method [14].

2.4. Evaluation

The models are validated on an independent test set. In our application, the test set consists of all records of 30% of randomly selected patient. As with the training set, a separate test set is created for each day d. An important performance aspect of a probabilistic model is its calibration ability, that is, the prediction of "faithful" probabilities, which are close to the true (unknown) probabilities of an event. To take this aspect into account we will apply the commonly used Brier score defined for two classes as:

$$\frac{1}{N}\sum_{i=1}^{N}(P(Y_i = 1|\mathbf{x}_i) - y_i)^2 \quad (1)$$

where N denotes the number of patients, and y_i denotes the actual outcome for patient i. The vector $\mathbf{x}_i$ represents the covariate values for patient i. The Brier score is a measure of error. It is a *strictly proper scoring rule* [15] which means it is optimal (lowest value) only when the true probability of the event is provided. In contrast, the area under the (ROC) curve (the AUC) is not *strictly* proper scoring rule meaning that its optimal value (highest AUC value) can be obtained in cases with probabilities different than the true ones [15,16]. The AUC only considers the ranking of subjects with or without the event and will not penalize models that under- or over-predict the probability as long as the relative order between subject remain the same. As a simple example, the AUC obtained by a set of predictions will be equal to that obtained by squaring these predictions. In this sense the AUC can make models "look" better than they actually are in terms of providing the true probabilities. This is not the case when a proper scoring rule, like Brier, is used. The performance of each of the five temporal models is compared to its corresponding static model—the model based only on SAPS—on the same corresponding test set. To test whether the difference between Brier scores of any static-temporal model pair is statistically significant on the corresponding test set we used the bootstrap method [17] with 1000 bootstrap samples. This is a non-parametric method that does not make distributional assumptions about the parameter under investigation.

3. Results

3.1. Frequent aligned episodes

Mining the data for the first five days resulted in five sets of frequent aligned episodes with the following sizes: 3, 8, 15, 14 and 19, respectively. Table 3 exemplifies the frequent episodes discovered for the 5th day (using data of the first 5 days of patients staying at least 5 days in the ICU). In addition, for each episode we show its support (in terms of frequency rate) in the data and the mortality rate of patients having the episode (aligned at the fifth day).

Table 3
Discovered frequent episodes for day 5

Episode	Support (%)	Mortality rate (%)
L	26	13
M	25	32
H	49	43
LL	18	10
ML	7	17
MM	16	33
MH	5	46
HM	7	29
HH	42	42
LLL	12	13
MMM	11	38
HHM	6	26
HHH	38	43
LLLL	10	14
MMMM	6	36
HHHH	35	45
LLLLL	7	10
MMMMM	5	35
HHHHH	34	45

3.2. Model development

Table 4 shows the five obtained static and temporal models, one for each day. The models are described by their logit where the dummy variables are denoted by the episode they represent. For example the logistic regression temporal model for day 3 is:

$$p(Y = 1|\text{SAPS}, \texttt{L}, \texttt{HHM}, \texttt{HHH}) = \frac{e^{-3+0.04\cdot\text{SAPS}-0.7\cdot\texttt{L}-1.5\cdot\texttt{HHM}+0.5\cdot\texttt{HHH}}}{1+e^{-3+0.04\cdot\text{SAPS}-0.7\cdot\texttt{L}-1.5\cdot\texttt{HHM}+0.5\cdot\texttt{HHH}}}$$

Using this formula, the probability of death for a patient with a SAPS score of say, 40, and the episode `HHH` is 0.29. For a patient with a SAPS of 40 but with the `HHM` episode the probability of death is merely 0.052. A patient can only have one of the episodes `L`, `HHM`, or `HHH`, because they imply a qSOFA on day 3 of `L`, `M`, or `H`, respectively, and the patient must have one of these mutually exclusive values.

Table 4
The temporal and static models, described by their logit

Day	Temporal model logit	Static model logit
1	$-6.3 + 0.1\cdot\text{SAPS} + 0.1\cdot\texttt{M} + 0.6\cdot\texttt{H}$	$-6.4 + 0.1\cdot\text{SAPS}$
2	$-4 + 0.06\cdot\text{SAPS} - 0.5\cdot\texttt{L} - 0.8\cdot\texttt{HM} + 0.3\cdot\texttt{H}$	$-4.4 + 0.06\cdot\text{SAPS}$
3	$-3 + 0.04\cdot\text{SAPS} - 0.7\cdot\texttt{L} - 1.5\cdot\texttt{HHM} + 0.5\cdot\texttt{HHH}$	$-3.5 + 0.05\cdot\text{SAPS}$
4	$-3.2 + 0.03\cdot\text{SAPS} + 1\cdot\texttt{M} + 1.1\cdot\texttt{H}$	$-2.7 + 0.04\cdot\text{SAPS}$
5	$-2.7 + 0.02\cdot\text{SAPS} + 1.1\cdot\texttt{M} + 1.3\cdot\texttt{H}$	$-2.2 + 0.03\cdot\text{SAPS}$

Table 5
Description of the covariates: SAPS and the frequent episodes

Day	Covariate	Support %	Deaths in episode %	Sample size #patients	Death %	Odds ratio
1	H	23	33	4389	11	1.8
	M	48	5.3			1.1
	SAPS	—	—			1.1
2	L	34	9	1236	23	0.6
	H	36	41			1.35
	HM	8	16			0.45
	SAPS	—	—			1.06
3	L	28	13	791	30	0.5
	HHM	6	8			0.22
	HHH	35	49			1.65
	SAPS	—	—			1.04
4	M	28	32	578	31	2.72
	H	46	41			3
	SAPS	—	—			1.03
5	M	25	32	444	32	3
	H	49	43			3.67
	SAPS	—	—			1.02

Table 6
Brier scores: temporal vs. static models

Day	Brier		Win
	Temp	Static	
1	0.058	0.059	Yes
2	0.128	0.132	Yes*
3	0.161	0.170	Yes
4	0.171	0.180	Yes*
5	0.166	0.182	Yes*

Table 5 describes the episodes that were selected in the models (appearing in Table 4) and the training datasets corresponding to each of the first five days. The table reports the data support of an episode, the mortality rate in patients having it, and the odds ratio (see Appendix) for those having the episode in comparison to those not having it.

3.3. Model evaluation

Table 6 shows the Brier scores obtained on the corresponding independent test sets for each of the models. It also reports on whether the models developed using frequent episodes won from the SAPS model. An asterisk (*) indicates that the difference in the Brier scores is statistically significant at the 0.05 level. It is apparent that all temporal models outperform the reference models based on SAPS alone.

4. Discussion and related work

In this paper we suggested and applied a new method for exploiting temporal information in Intensive Care prognosis. The novelty of the method stems from adapting and integrating an existing technique for mining frequent temporal episodes within probabilistic predictive modeling. The integration is achieved by representing frequent episodes in terms of dummy variables in a logistic regression model that is obtained by using a model selection strategy based on an information-theoretic measure and avoidance of logically entailed collinear covariates. Our real life case study demonstrated the added value of these episodes in survival state predictions of Intensive Care patients by generating more accurate models. We did not evaluate the applicability of these results in terms of the clinical or managerial tasks that can be supported such as adapting treatment or workload scheduling. These form new research topics to be investigated. The method can be applied in various probabilistic prediction problems in temporal domains. However, in any application of the method, various choices are to be made in order to tailor it to the problem at hand. Below we discuss the results of our method followed by a discussion of our approach.

4.1. Discussion of results

Analysis of the selected discovered frequent episodes, like those shown in Table 3 for day 5, provide insight into their relation to mortality and into the patient population. Consider the episode ML. It occurred in 7% of the patients, meaning that their qSOFA improved from M on the fourth day to L on the fifth day. Only 17% of these patients eventually died. Compare this with the mortality rate of 45% among patients with H on all five days.

The longer frequent episodes (length 3 on) seem to mostly describe constant trends like HHHH. One reason for the emergence of such episodes is the choice for only three categories. This results in their relative high frequency especially in categories covering a large range of values. For example, a decrease from a SOFA score of 13 to a value of 12 or 10 on the following day will belong in either case to the episode HH. Allowing for more categories would capture smaller changes from one day's score to another at the risk of compromising support.

We also note that patients staying a relatively long number of days in the ICU are associated with higher SOFA scores. Within these patients, the support of episodes like HHHH is higher than the other equally long episodes. Intuitively the relatively constant trend in mortality ass1ociated with the logically entailed episodes (e.g. L, LL, LLL, LLLL, LLLLL) is due to the fact that there is a big overlap between the respective patient groups and because the last days are the most associated with mortality. The effect of the shorter episodes on prediction will be dominant due to the relatively large number of patients associated with them.

Our predictive models creation approach makes use of a variable selection strategy based on an information-theoretic criterion. The obtained results depicted in Table 4 call upon some explanatory remarks. A first observation is that only a very small fraction of frequent episodes have been selected in the models. For example out of the 19 frequent episodes of day 5 which appear in Table 3 only M and H have been retained. A second observation is that

elementary episodes (L,M, and H) are common in the models appearing in Table 4. This observation is partially due to our decision not to allow for collinear covariates. This means for example that while the episodes LH and MH might coexist in a model, the episodes H and MH cannot, as MH logically entails H.

An advantage of using a logistic regression model is the interpretability of its coefficients and their associated odds ratios. The odds ratios presented in Table 5 are in concordance with clinical knowledge. Episodes indicating low scores or a decrease in SOFA scores correspond to odds ratios <1 and contributing to a higher probability of survival. For example, the episode HHM selected in the model for day 3 occurred in 6% of the patients that stayed at least 3 days (791 patients with a mean mortality of 30%). Only 8% of the patients having this episode died. The odds ratio of 0.22 (calculated from the model in Table 4 by $e^{-1.5}$) means that the odds for dying for patients with this episode is only 0.22 times of that for patients not having this episode (alternatively, the latter patients have 4.5 times the odds of dying compared to patients who had the episode). For episodes indicating high scores, like HHHH, the odds ratios are >1 reinforcing the belief that they diminish the survival chances.

Table 6 showed the performance of both model types, temporal and static, measured by the Brier score. Smaller values of the Brier score mean better accuracy, which includes elements of discrimination as well as calibration (see [15]). Models, in our case the temporal models, cannot be ''proven'' correct but one hopes to get incremental evidence for their validity or superiority. In our case the evidence for the validity and superiority of the temporal models consists of the following. First, the selected covariates in the model, using the AIC criterion, included the episodes. Second, on an independent test set, the temporal models performed better than the static ones in every of the 5 days, although one should not forget that the data on later days is dependent on data of earlier days. Third, based on bootstrapping, in 3 out of the 5 days the difference between the Brier scores turned out to be statistically significant (denoted by the asterisk (''*'')). Recall that a statistical significant test measures the improbability of an obtained statistic (e.g. difference between Brier scores) assuming that the null hypothesis (e.g. that the models performances are equal) is true. Not rejecting the null-hypothesis should not be interpreted as accepting it. To summarize, there is ample evidence that the temporal models learnt on the *given training set* are superior to the static ones when tested on the *given test set*. This is hence no claim to the superiority of our method in general. Such claims require a different design as described later on in this section.

4.2. Discussion of approach

Below we discuss the merits and limitations of the choices made in our approach and illustrate them in the context of our application in the Intensive Care.

4.2.1. Categorization

When dealing with continuous or integer valued variables one should decide on the number of categories and the categorization method. Too many categories would not have enough support in the data. Too few ones could blur important distinction between values. This tradeoff is discussed in the context of association rules in [18] where incremental joining of adjacent intervals is suggested. We have chosen for only three categories in our application and did not consider creating new categories of adjacent basic categories (e.g. the category encompassing the L and M categories). Further work will investigate the merits of using more categories and joining them adaptively as discussed in [18].

We have chosen for equally sized quantiles of SOFA scores for categorization as suggested in [18]. Quantilization forces the proportion of the most frequent values to reside in different categories. This mitigates the domination in the episodes of categories associated with frequently occurring values in a narrow interval. For example if 95% of the SOFA scores were in the interval [9,10] then there will only be one frequent category. However the resulting categories from quantilization are not necessarily meaningful nor guaranteed to be the most useful in prediction. Further work consists of investigating clinically meaningful categories and other ways for categorization. One such categorization method that takes the outcome into consideration is based on the entropy of the outcome that the categories imply as described in [19]. A practical way to do this is to take a summary measure for the SOFA scores of each patient (e.g. the maximum) and fit an entropy-based classification tree for the outcome.

4.2.2. Discovery of frequent episodes

Using only frequent episodes for model development has several advantages. First, they are representative of the relatively significant subpopulations of patients. Second, they strongly simplify the model selection process because the non-frequent episodes, which comprise the majority of episodes, are unlikely to be selected in the model. Indeed, the selected episodes in the models in our applications tend to have relatively high support. However, future research should investigate the inclusion of evaluation measures pertaining to the task at hand, here prediction, other than frequency, especially if one is seeking interesting patient subgroups, instead of seeking a global predictive model. One way to go about this is the extraction of high confidence Association Rules having mortality as the sole variable on the right-hand side of each rule.

Our choice for episodes which are sequential temporal episodes aligned to the day of prediction makes them easy to interpret and does not burden the modeler with prespecifying various types of episodes. However, more expressive episodes and those motivated by the domain at hand should be further explored. For example one could allow for sequential episodes that solely provide the relative order between the qSOFA scores without constraining the scores to correspond to consecutive days. An example

of a clinically motivated episode type is one in which also trends can be expressed.

Considering the SOFA sub-scores for six individual organ systems in the analysis seems a natural extension to our method. One could use the sub-scores separately or seek a custom combination thereof by e.g. using a weighted summation of their values where higher weights are given to the most predictive score. The predictive capabilities of each organ system score could be assessed with a univariate analysis. Other clinically intuitive abstraction are also possible such as the inclusion of the number of organ failures on the day of prediction or its mean in the last days prior to prediction.

4.2.3. Model development

Using logistic regression as the formalism of choice for prediction has the advantage of employing an established and well understood framework. The coefficients of the model have a meaningful interpretation. The use of Akaike's Information Criterion (AIC) for variable elimination overcomes several drawbacks of methods depending on p-values and at the same time penalizes the models for their complexity in terms of the number of selected covariates. In the evaluation of these models we only used the split-sample design and used the Brier score. This provides evidence that temporal models are superior to the static models for this test set. This is however no claim for the general superiority of the temporal models on other test sets, and indeed not a claim for the general superiority of the learning algorithm. The algorithmic evaluation of our method consists of further work that we are currently undertaking. This evaluation requires a different design in which multiple models, not just one, are created on more days and evaluated on various test sets with different performance measures. This evaluation also requires more data than used in this work.

4.2.4. Intensive Care literature

The relatively few related work in the Intensive Care literature that investigates the SOFA scores in the context of mortality prediction is focused on showing a positive correlation between the SOFA scores and mortality, e.g. [6]. This is usually done by showing that non-survivors have significantly higher SOFA scores than survivors in different patient subpopulations. Work which does explicitly include abstractions of SOFA scores would usually employ simple statistics, as typified by [20,21], such as the Δ (difference between admission score and maximum score during stay) or maximum score at admission or at the last day of stay. There is some work that investigates pre-specified temporal trends defined as changes over time, such as "increase" or "decrease". Such work is limited to considering a very short number of days (e.g. first 2–4 days) [22–24]. Our work is different in the sense that it uses data-driven discovery of episodes. Moreover it investigates the added value of frequent episodes in relation to the static information.

4.2.5. Data Mining literature

Related work in Data Mining focusses on computational and complexity aspects of algorithms for the discovery of patterns [25] and using them in various forms of association rules. Other uses of patterns includes clustering them in a hierarchy as described in e.g. [26]. Our discovery algorithm is based on that described in [7] except that episodes are discovered for a specific day, the day of prediction, and in addition they are aligned to it. This means that we do not assume a stationary process generating the time series. From a data mining perspective the novelty of the proposed method is showing a new way to use patterns in predictions, beyond their traditional use in association rules.

Perhaps the most similar works to ours, in their general aim and application, are those described in [27] and [28]. In [27] patterns are discovered from multivariate series in order to predict mortality in the Intensive Care in six consecutive periods in the future. That work is different in various ways than ours in terms of the assumption of stationarity (see [29] for discussion) of the time series, the non Apriori-like generation of patterns, the assessment of patterns based on their discriminative ability (in terms of the area under the ROC curve). Like in [27], our episodes have various lengths and in this sense non stationarity is assumed. However we also allow for a second source of non-stationarity in our work by requiring alignment of the episodes to the day on which prediction is made, this means for example that the frequency of episodes relies on the positioning of the episode in the time-series. More important differences reside however in three other elements. First, the work of Kayaalp et al. does not use a standard static model, as in our case, in order to assess the added value of patterns. Second, it uses a patient-specific model: only patterns identified in a patient test case are ranked and the best ones are used for prediction. The other discovered patterns in the training set are not used in the ranking. This is an interesting idea that we did not explore in our work. Third, the patient-specific selected patterns are combined using the Naive Bayes Classifier. To provide probability estimates, the Naive Bayes approach assumes the independence of patterns conditioned on the outcome (e.g. $P(ptt1,ptt2|Y=1) = P(ptt1|Y=1)\cdot P(ptt2|Y=1)$). It is demonstrated in the machine learning literature that this assumption is often (mildly) violated but that it usually results in adequate *classification* models (e.g. to survival or death). However if one is interested in the exact probabilities of the event, as we aim at in our work, and not just the predicted events themselves, the conditional independence assumption can lead to poor probability estimates. Poor estimates will also go undetected when a non-proper scoring rule, such as the AUC, is used. In our specific case of using aligned episodes, the (conditional) independence assumption would be severely violated due to the presence of logically entailed episodes: they are clearly very much dependent on each other. Our use of episodes inside logistic regression does not only allow for integrating new methods into the established framework in IC prediction, but also

takes into account the inter-dependency between all covariates.

In [28] four temporal measurements of variables, some of which are also used for calculating the SOFA scores, are summarized to obtain what is referred to as adverse events which correspond to extreme values of the variables and can span over a relatively long period of time. A number of 12 new variables are constructed, 8 describing the daily average number of events and critical events (events with larger duration or even more extreme values of their measurement) and 4 describing the daily averages time span, in minutes, of each of the 4 types of critical events. The adverse events information is fed into a neural network (NN) and a logistic regression model for predicting ICU mortality. This information consists of the logarithmic transformations of the daily averages of the number of the events' occurrences and durations. The SAPS-II model was included in the analysis and the SAPS-II score was a covariate in some setups together with the adverse events. The NN and logistic regression using temporal information were shown to have better performance in terms of discrimination ability compared to the other models, measured as the AUC. The approach followed in [28] assumes a much stricter form of stationarity since the mean value of adverse events is taken to characterize the process. This approach hence disregards the measurement time and abstracts away from the explicit temporal evolution in the data. In contrast, capturing this evolution comprised one of the starting points in our work.

In summary, our aim in this paper was the development of a method that captures the temporal evolution of organ functioning and, at the same time, embeds it in the current logistic regression modeling framework. We expected that the integration of temporal data will bear fruit as increasing the predictions' accuracy. We attain this goal by proposing, and applying a method for integrating a data-driven approach for mining of frequent patterns (called episodes) in the current logistic regression framework for probabilistic predictions. Our results in Intensive Care mortality prognosis can be regarded as a proof of concept of the merits of this method. The noticed concordance with the results of similar works (e.g. in [27,28]) is another reinforcing reason for research on the usage of temporal data for prediction. Further work consists of investigating the effects of alternative choices or tuning of settings for categorization, type and quality of episodes, model fitting strategy, and performance measures. We are currently conducting a comparative evaluation study in our IC application domain addressing the questions of when and whether the *method*, not just the specific models developed in this paper, is superior to the method that trains models to each day of prediction but does not use temporal patterns.

Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research (NWO) under the I-Catcher project number 634.000.020. We thank Arno Siebes, Niels Peek, Barry Nannings, Linda Peelen, Marion Verduijn, Frans Voorbraak, and Floris Wiesman, for their helpful discussions on the topics described in this paper.

Appendix A. Logistic regression

A logistic regression model [30] is a parametric model specifying the conditional probability of a binary outcome variable Y, given the values of the covariates of the model. In our case, $Y = 1$ indicates the occurrence of a death event. The logistic model has the following form:

$$p(Y = 1|\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \tag{2}$$

where $\mathbf{x} = (x_1, \ldots, x_m)$ is the covariate vector. For m covariates (also called predictors, input or independent variables) the *logit function* which is defined as $\log(\frac{p(Y=1|\mathbf{x})}{1-p(Y=1|\mathbf{x})})$ is equal to $g(\mathbf{x})$ which has the following linear form:

$$g(\mathbf{x}) = \beta_0 + \sum_{i=1}^{m} \beta_i \cdot x_i \tag{3}$$

where β_i, $i = 1, \ldots, m$, denote the coefficients of the m covariates. In the temporal models in this paper, one of the covariates is the SAPS, the other covariates are dummy variables denoting frequent episodes. One reason for the popularity of the model is the interpretation that is given to each β_i in terms of an *odds ratio*. Suppose the logit function is $\beta_0 + \beta_1 \cdot SAPS + \beta_2 \cdot Ep$ where $Ep = 1$ for patients having some specific episode and *0* for patients not having the episode. The odds of dying for those having the episode, $odds(Ep = 1)$, is $P(Y = 1|Ep = 1)/P(Y = 0|Ep = 1)$ and for those not having the episode, $odds(Ep = 0)$, is $P(Y = 1|Ep = 0)/P(Y = 0|Ep = 0)$. The quantity e^{β_2} turns out to be equal to the odds ratio $odds(Ep = 1)/odds(Ep = 0)$. If there is no difference between the odds for those with the episode and those without it, assuming all other variables (in this case only SAPS) have the same values, the odds ratio will be 1. A higher value indicates higher risk to die for those having the episode, and a lower value than 1 indicates higher risk for those who do not have it. The interpretation of e^{β_1} is similar, it indicates the odds ratio between a group of patients who have a SAPS of one unit more than the other group, averaged on those with or without the episode denoted by Ep.

References

[1] Abu-Hanna A, Lucas PJF. Prognostic models in medicine—AI and statistical approaches [editorial]. Methods Inform Med 2001;40:1–5.
[2] Le Gall J, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS-II) based on a European/North American multicenter study. J Am Med Assoc 1993;270:2957–63.
[3] Knaus W, Draper E, Wagner D, Zimmerman J. APACHE II: a severity of disease classification system. Crit Care Med 1985;13:818–29.
[4] Vincent JL, Ferreira FL. Evaluation of organ failure: we are making progress. Intensive Care Med 2000;26:1023–4.

[5] McClish DK, Powell SH. How well can physicians estimate mortality in a medical intensive care unit? Med Decis Making 1989;9:125–32.
[6] Vincent JL, de Mendonca A, Cantraine F, Moreno R, Takala J, Suter PM, et al. Use of the SOFA score to assess the incidence of organ dysfunction/failure in intensive care units: results of a multicenter, prospective study. Crit Care Med 1998;26(11):1793–800.
[7] Mannila H, Toivonen H, Verkamo AI. Discovery of frequent episodes in event sequences. Data Min Knowl Disc 1997;1(3):259–89.
[8] Shahar Y. A framework for knowledge-based temporal abstraction. Artif Intell 1997;90:79–133.
[9] Bellazzi R, Larizza C, Magni P, Bellazzi R. Temporal data mining for the quality assessment of hemodialysis services. Artif Intell Med 2005;34:25–39.
[10] Agrawal R, Srikant S. Fast algorithms for mining association rules. In: Bocca JB, Jarke M, Zaniolo C, editors. Proceedings of the 20th very large data bases conference. Morgan Kaufmann; 1994. p. 487–99.
[11] Harrell Jr FE. Regression modeling strategies. Springer Series in Statistics; 2001.
[12] Burnham KP, Anderson DR. Model selection and multimodel inference: a practical-theoretic approach. New York: Springer; 2002.
[13] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning data mining, inference, and prediction. Springer Series in Statistics; 2001.
[14] Venables WN, Ripley BD. Modern applied statistics with S. Springer Series in Statistics and Computing; 2003.
[15] Hand JD. Construction and assessment of classification rules. John Wiley & Sons; 1997.
[16] Abu-Hanna A, Keizer NF. Integrating classification trees with local logistic regression in intensive care prognosis. Artif Intell Med 2003;29(1–2):5–23.
[17] Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall; 1993.
[18] Srikant R, Agrawal R. Mining quantitative association rules in large relational tables. In: Jagadish HV, Mumick IS, editors. Proceedings of the international conference on management of data. ACM Press; 1996. p. 1–12.
[19] Kohavi R, Sahami M. Error-based and entropy-based discretization of continuous features. In: Simoudis E, Han J, Fayyad U, editors. Proceedings of the 2nd international conference on knowledge discovery and data mining. AAAI Press; 1996. p. 114–9.
[20] Kajdacsy-Balla Amaral AC, Andrade FM, Moreno R, Artigas A, Cantraine F, Vincent JL. Use of sequential organ failure assessment score as a severity score. J Intensive Care Med 2005;31:243–9.
[21] Moreno R, Vincent JL, Matos R, Mendonca A, Cantraine F, Thijs L, et al. The use of a maximum SOFA score to quantify organ dysfunction/failure in intensive care. Results of a prospective, multicentre study. J Intensive Care Med 1999;25:686–96.
[22] Levy MM, Macias WL, Vincent JL, Russell JA, Silva E, Trzaskoma B, et al. Early changes in organ function predict eventual survival in severe sepsis. Crit Care Med 2005;33(10):20–2194.
[23] Ferreira FL, Bota DP, Bross A, Melot C, Vincent JL. Serial evaluation of the SOFA score to predict outcome in critically Ill patients. J Am Med Assoc 2001;286:1754–8.
[24] Cabre L, Mancebo J, Solsona JF, Saura P, Gich I, Blanch L, et al. Multicenter study of the multiple organ dysfunction syndrome in intensive care units: the usefulness of sequential organ failure assessment scores in decision making. J Intensive Care Med 2005;31(7):927–33.
[25] Pei J, Han J, Mortazavi-Asl B, Wang J, Pinto H, Chen Q, et al. Mining sequential patterns by pattern-growth: the prefixspan approach. IEEE Trans Knowl Data Eng 2004;16(10):1424–40.
[26] Bathoorn R, Siebes A. Constructing (almost) phylogenetic trees from developmental sequences data. In: Boulicaut J-F, Esposito F, Giannotti F, Pedreschi D, editors. Proceedings of the 8th European conference on principles and practice of knowledge discovery in databases. Springer-Verlag; 2004. p. 500–2.
[27] Kayaalp M, Cooper G, Clermont G. Predicting with variables constructed from temporal sequences. In: Jaakkola T, Richardson T, editors. Proceedings of the 8th international workshop on artificial intelligence and statistics. Morgan Kaufmann; 2001, p. 220–5.
[28] Silva A, Cortez P, Santos MF, Gomes L, Neves J. Mortality assessment in intensive care units via adverse events using artificial neural networks. J Artif Intell Med 2006;36:223–34.
[29] Kayaalp M, Cooper G, Clermont G. Predicting ICU mortality: a comparison of stationary and nonstationary temporal models. In: Overhage J, editor. Proceedings of the American medical informatics association symposium; 2000, p. 418–22.
[30] Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons; 1989.